University of Sheffield, Electronic Systems Group Report No. 95/44
Also appears in: D.G.Storck (Editor), NATO ASI Speechreading by Man and Machine: Models, Systems and Applications, Springer Verlag, 1996.

Active Shape Models for Visual Speech Feature Extraction

Juergen Luettin, Neil A. Thacker, Steve W. Beet
University of Sheffield, UK

Abstract. Most approaches for lip modelling are based on heuristic constraints imposed by the user. We describe the use of Active Shape Models for extracting visual speech features for use by automatic speechreading systems, where the deformation of the lip model as well as image search is based on *a priori* knowledge learned from a training set. We demonstrate the robustness and accuracy of the technique for locating and tracking lips on a database consisting of a broad variety of talkers and lighting conditions.

Keywords. lip locating, lip tracking, learned model, learned features

1. Introduction

While mainstream speech recognition research has concentrated almost exclusively on the acoustic speech signal, it is well known that humans use visual information of the talker's face (mainly lip movements) in addition to the acoustic signal for speech perception purpose. Whereas several well known methods exist for representing acoustic features of speech, it is still not fully understood (i) which visual features are important for speechreading, (ii) how to extract them and (iii) how to combine them with the acoustic information. It is generally agreed that most visual information is contained in the lips, especially the inner lip contours and to a minor extent in the visibility of teeth and tongue (Montgomery and Jackson 1983, Summerfield 1992). The main difficulty in incorporating information about lip movements into an acoustic speech recognition system is to find a robust and accurate method for extracting important visual speech features. The technique should be able to locate and track lips in faces of various talkers and should be robust to lighting, rotation, scale and translation (LRST). The extracted features should be sensitive to variances which account for different visemes and insensitive to variances which account for linguistic variability and image variability (LRST).

Here we describe the use of Active Shape Models (ASMs), introduced by Cootes et al. (1994), for robust detection, tracking and parameterisation of visual speech information. In comparison to previous contour tracking approaches such as the use of deformable templates (Yuille, Hallinan and Cohen 1992, Hennecke, Prasad and Stork 1994) or snakes (Kass, Witkin and Terzopoulos 1988, Bregler and Omohundro 1994), ASM is a statistically based technique which almost completely avoids the use of constraints, thresholds or penalties imposed by the user. During image search, the model is only allowed to deform to shapes similar to the ones seen in the training set. Whereas deformable templates and snakes align to strong gradients for locating the object, regardless of their actual appearance in the image, ASMs learn the typical grey level appearance perpendicular to the contour from the training set and use them for image search.

2. Active Shape Models

Active Shape Models are flexible models which represent an object by a set of labelled points. The points describe the boundary or other significant locations of an object. Figure 1 shows one of the models we used for the lips. Although *s*, represented by the width of the lips, does not characterise the real scale of the lip model, we used it to enable us to map different examples of lip shapes to each other. We built two different models of the lips: Model 1 describes the outer contour of the lips and Model 2 describes the outer and inner contour of the lips.

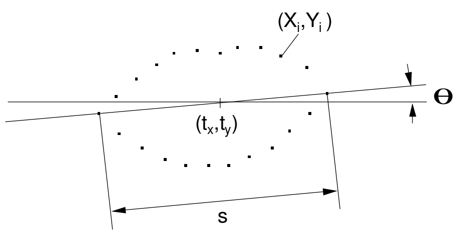

Figure 1: Model 1, representing the outer lip contour.

2.1 Capturing shape statistics

ASMs make no heuristic assumptions about legal shape deformation. Instead *a priori* knowledge about typical deformation is obtained by examining a training set in order to obtain the average shape and the main modes of variation. The first training set is labelled by hand to build an initial model. This model can then be used for "boot-strapping" further training sets. The training shapes are normalised by scaling to unit width, zero translation and zero rotation.

Because the points are evenly spaced along the horizontal, for further processing we only need to regard the vertical parameter of each point to describe a desired shape. Principal component analysis (PCA) is performed on the training shapes to obtain the main modes of variation. Any normalised shape

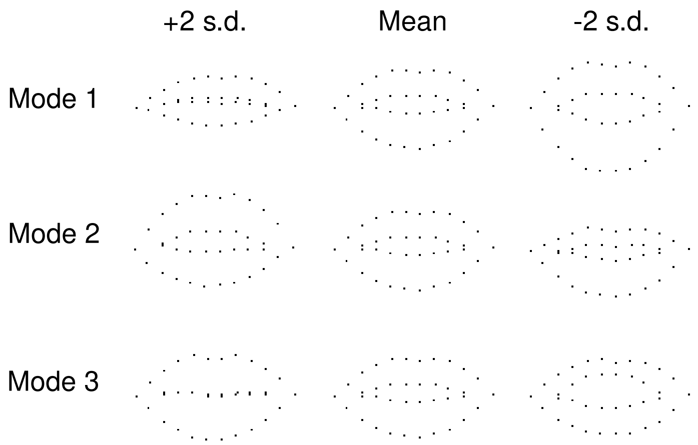

Figure 2: Mean shape and main modes of variation of Model 2.

can then be approximated using the mean shape and the first few principal modes of variation. Figure 2 shows the mean shape of Model 2 and the first three modes of variation by +/- 2 standard deviation (s.d.). This approach assumes that no non-linear dependencies are present between the modes.

2.2 Capturing grey level statistics

In order to use ASMs for image search, a cost function is needed which measures the fit between the model and the image. We therefore need to find a way of representing dominant image features of the lip contours. The most common approach for representing contours is to use edges or gradients. However, the lip contour can appear in many different ways. The gradient has different values along the contour and is dependent on talker, illumination, reflections, visibility of teeth and mouth opening. Figure 3 shows two examples from the database we used and their gradient magnitude images after Gaussian smoothing. The examples show clearly the difficulties gradient-based search methods are faced with and that gradient information is not an appropriate way to represent dominant features of lip contours.

In analogy to the statistical description of the lip deformation we want to avoid the use of heuristics for image search and rather use learned knowledge of the actual grey level appearance at the contour as well as it's global variation across different examples. Assuming that grey-level changes are not only important at each contour point but also in regions around each point, we capture the statistics of the actual grey level appearance around each model point and estimate their main modes of variation within a training set.

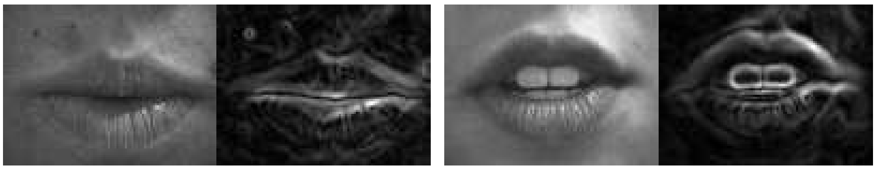

Figure 3: Example images and their gradient images.

Following Cootes et al. (1994) we choose to sample one-dimensional profiles perpendicular to the contour and centred at the model points for each training image. But instead of calculating individual mean profiles and covariance matrices for each model point, we concatenate the profiles of all model points to construct a global profile vector for each training image. The principal modes of profile variation are obtained by performing PCA on the global training profiles. Any profile of the training set can then be approximated using the mean profile and the first few main modes of profile variation. This approach assumes that the deviations of profiles at different model points are correlated with each other, as they are expected to be due to illumination effects and differing skin colour.

2.3 Locating lips

For locating the lips in the first frame of an image sequence we assume that a rough estimate of the region of interest (ROI) containing the lips is known from another image processing algorithm. The model is initialised with the mean shape and placed in a random location in the ROI.

Instead of using the search technique described in Cootes et al. (1994) we use a cost function to evaluate how well a model fits the image and the Downhill Simplex Method, introduced by Nelder and Mead (1965), for finding a minimum. The algorithm was implemented as described in Press et al. (1992). To measure how well a model fits the image, the cost function measures the grey-level distance between the model profile and the image profile. To account for grey level variation captured in the training set, the model profile is aligned to the image profile as closely as possible using the mean profile and the first few main modes of variation.

To restrict the model to only deform to shapes similar to the ones seen in the training set we constrain the shape parameters for each mode to stay within +/- 3 s.d. (+/- 3 s.d. account for 99% of variance assuming a Gaussian probability density function). Similarly we restrict the profile parameters for each mode to stay within +/- 3 s.d.. In comparison to other methods, no internal energy component of the model is added to the cost, assuming equal prior probabilities for all model shapes within the limits of +/- 3 s.d. Examples of locating lips are shown in Figure 4.

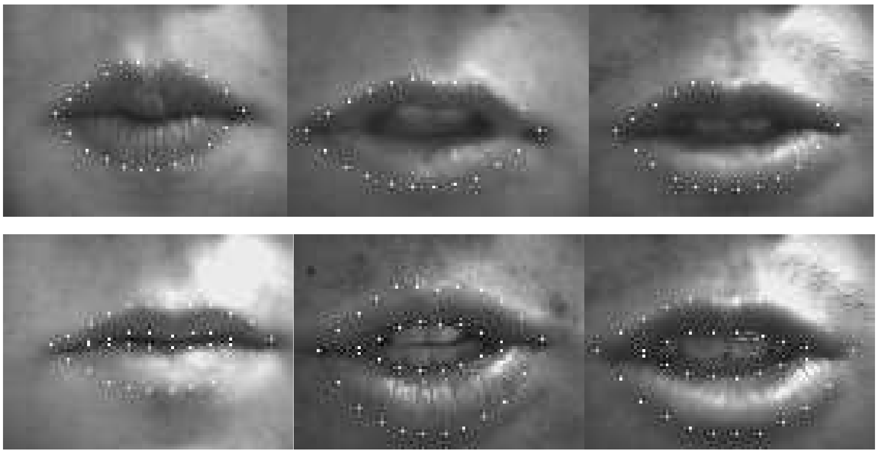

Figure 4: Examples of lip location with Model 1 (top row) and Model 2 (bottom row) with off centre initialisation.

2.4 Tracking lips

For lip tracking, the lips are first located in the first frame as described above. For consecutive frames the previous frame is used as the initial estimate of the lip position and the search is performed using the same search algorithm. Although constraints could be introduced to limit the search to stay within certain limits during tracking, for simplicity we used the same constraints as for locating the lips. Figure 5 shows tracking examples for Model 1 and Model 2.

The shape parameters and scale information extracted from an image sequence can directly serve as visual speech features for a speech recognition system. The shape parameters are invariant to scale, translation, rotation and lighting.

3. Experiments

We used the database described in Movellan 94 for our experiments. It consists of 96 image sequences of 12 talkers (9 male, 3 female) each saying the first four digits in English twice (later described as set 1 and set 2). It contains images of a variety of talkers with different skin and lip colour, lip shape and illumination.

The shape and profile models were built using the first frame of each sequence of set 1 with a total of 48 images. We used 6 shape modes for Model 1, which covers 81% of shape variation and 8 shape modes for Model 2, which covers 76.9% of shape variation. All tests for locating the lips were performed on set 2. For location tests the model was initialised at a position off the centre as seen in

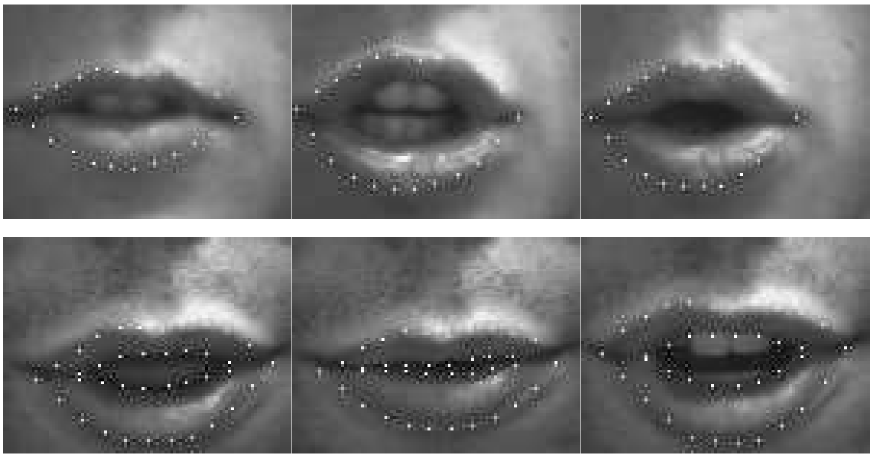

Figure 5: Examples of lip tracking with Model 1 (top row) and Model 2 (bottom row).

Figure 6, to make the search process more realistic. For tracking tests the model was initialised at the centre of the image.

For comparison we also performed tests by using the gradient magnitude of the image instead of the profile model. The images were first Gaussian filtered before calculating the gradient magnitude. The resulting images were smoothed with an exponential function to blur the gradients over a large image area. The cost was defined as the negative sum of the gradient image of all model points. Since the results for locating the lips were so poor, we performed another test where the model was initialised at the centre of the image.

4. Results

The image search was judged by visual inspection. A search result was classified as *Good* if the whole lip contour was found within one quarter of the lip thickness deviation, it was classified as *Adequate* if it was found within half the lip thickness deviation and it was classified as a *Miss* otherwise.

Experiments with different numbers of profile modes showed that best results were achieved by using 6 profile modes for Model 1 and 12 profile modes for Model 2. Table 1 shows the results for lip location using profile models and different numbers of profile modes. The accuracy generally increased by adding more modes up to a certain point when the accuracy decreased again. We believe that this is due to the small set of training examples which results in inaccurate estimations of minor modes of variation.

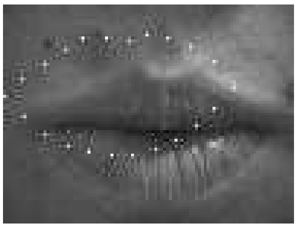

Figure 6: Off centre initialisation of the model.

Table 2 shows the results for locating the lips by using gradient information instead of the profile model. Mis-location was mainly caused by the weak gradient of the lower lips, reflections on the lips and gradients originating from the teeth. Initialising the model in the centre of the image improved the results only slightly.

Table 3 shows the results for tracking the lips. For Model 1, results for tracking the lips are similar to the ones for location. However, for Model 2, the performance for tracking is lower than for locating. Most of the *Adequate* and *Miss* results were caused by the inner lip contour of the model being unable to follow the image contour. We believe that this is due to the small training set which consisted of only the first image of each sequence and therefore did not contain many images with the mouth being open. We also believe that the very small training set is not sufficient for describing a high dimensional profile covariance matrix. Since the profiles inside the inner lip contour mainly change between three different levels (lips closed; mouth open and teeth visible; mouth open and teeth not visible) it is unlikely that a Gaussian distribution is appropriate to describe this variation.

Table 1: Results for lip location using profile models with off-centre initialisation.

	Profile Modes	Good (%)	Adequate (%)	Miss (%)
Model 1	0	72.9	25	2.1
Model 1	6	89.6	8.3	2.1
Model 2	0	72.9	18.8	8.3
Model 2	12	91.7	6.2	2.1

Table 2: Results for lip location using gradient search.

	Initialisation	Good (%)	Adequate (%)	Miss (%)
Model 1	off centre	6.3	16.7	77
Model 2	off centre	4.2	10.4	87.5
Model 1	centre	8.3	18.8	72.9
Model 2	centre	14.6	14.6	70.8

Table 3: Results for lip tracking.

	Good (%)	Adequate (%)	Miss (%)
Model 1	81.25	12.5	6.25
Model 2	62.5	20.8	16.7

5. Conclusions

We have shown that ASMs are able to represent deformable objects such as lips at various degrees of detail and with a small number of parameters. The deformation is purely governed by statistics learned from a training set and therefore neither too constrained nor too flexible.

The use of grey level profiles has been shown to be an appropriate way to represent dominant image features and to model their variation which enables accurate location of lips. The profile used in image search encompasses a large sample space which minimises the risk of terminating the search process at a local minimum.

ASM do not depend on heuristic thresholds, limits, weights or penalties imposed by the user. All limits used are related to probabilities derived from statistics of a training set.

Acknowledgements

Juergen Luettin is funded by a University of Sheffield Studentship and the German Academic Exchange Service (DAAD). The Authors would like to thank Dr. Movellan for making the speechreading database publicly available.

References

C.Bregler and S.Omohundro, “Surface Learning with Applications to Lip-Reading”, J.D.Cowan, G.Tesauro and J.Alspector (eds.) Advances in Neural Information Processing Systems 6. San Francisco, CA: Morgan Kaufmann Publishers, pp. 43-50, 1994.

T.F.Cootes, A.Hill, C.J.Taylor and J.Haslam, “Use of acrive shape models for locating structures in medical images”, Image and Vision Computing, Vol. 12, No. 6, pp. 355-365, 1994.

M.E.Hennecke, K.V.Prasad, D.G.Stork, “Using Deformable Templates to Infer Visual Speech Dynamics”, 28th Annual Asilomar Conference on Signals, Systems and Computers, pp. 578-582, 1994.

M.Kass, A.Witkin and D.Terzopoulos, “Snakes: active contour models”, Int. J. Computer Vision, pp. 321-331, 1988.

A.A.Montgomery, P.L.Jackson, “Physical characteristics of the lips underlying vowel lipreading performance”, J. Acoust. Soc. Am., Vol. 73, pp. 2134-2144.

J.R.Movellan, “Visual Speech Recognition with Stochastic Networks”, G.Tesauro, D.Touretzky, T.Leen (eds.) Advances in Neural Information Processing Systems. Volume 7, MIT Press Cambridge, pp. 851-858, 1995.

J.A.Nelder, R.Mead, “A simplex method for function minimization”, Comput. J. Vol. 7(4), pp. 308-313, 1965.

E.Petajan, “Automatic Lip Reading to Enhance Speech Recognition”, IEEE CVPR, pp. 44-47, 1985.

W.H.Press, S.A.Teukolsky, W.T.Vetterling, B.P.Flannery, “Numerical Recipes in C”, Cambridge University Press, Cambridge, 1992.

Q.Summerfield, “Lipreading and audio-visual speech perception”, Phil. Trans. R. Soc. Lond. B 335, pp. 71-78, 1992.

A.L.Yuille, P.Hallinan, D.S.Cohen, “Feature extraction from faces using deformable templates”, Int. J. Computer Vision, Vol. 8, pp. 99-112, August 1992.